

Edit Distance for Pushdown Automata*

Krishnendu Chatterjee¹, Thomas A. Henzinger¹, Rasmus Ibsen-Jensen¹, and Jan Otop²

¹IST Austria

²University of Wrocław

March 2, 2017

Abstract

The edit distance between two words w_1, w_2 is the minimal number of word operations (letter insertions, deletions, and substitutions) necessary to transform w_1 to w_2 . The edit distance generalizes to languages $\mathcal{L}_1, \mathcal{L}_2$, where the edit distance from \mathcal{L}_1 to \mathcal{L}_2 is the minimal number k such that for every word from \mathcal{L}_1 there exists a word in \mathcal{L}_2 with edit distance at most k . We study the edit distance computation problem between pushdown automata and their subclasses. The problem of computing edit distance to a pushdown automaton is undecidable, and in practice, the interesting question is to compute the edit distance from a pushdown automaton (the implementation, a standard model for programs with recursion) to a regular language (the specification). In this work, we present a complete picture of decidability and complexity for the following problems: (1) deciding whether, for a given threshold k , the edit distance from a pushdown automaton to a finite automaton is at most k , and (2) deciding whether the edit distance from a pushdown automaton to a finite automaton is finite.

1 Introduction

Edit distance. The edit distance [16] between two words is a well-studied metric, which is the minimum number of edit operations (insertion, deletion, or substitution of one letter by another) that transforms one word to another. The edit distance between a word w to a language \mathcal{L} is the minimal edit distance between w and words in \mathcal{L} . The edit distance between two languages \mathcal{L}_1 and \mathcal{L}_2 is the supremum over all words w in \mathcal{L}_1 of the edit distance between w and \mathcal{L}_2 .

Significance of edit distance. The notion of *edit distance* provides a quantitative measure of “how far apart” are (a) two words, (b) words from a language, and (c) two languages. It forms the basis for quantitatively comparing sequences, a problem that arises in many different areas, such as error-correcting codes, natural language processing, and computational biology. The notion of edit distance between languages forms the foundations of a quantitative approach to verification. The traditional qualitative verification (model checking) question is the *language inclusion* problem: given an implementation (source language) defined by an automaton \mathcal{A}_I and a specification (target language) defined by an automaton \mathcal{A}_S , decide whether the language $\mathcal{L}(\mathcal{A}_I)$ is included in the language $\mathcal{L}(\mathcal{A}_S)$ (i.e., $\mathcal{L}(\mathcal{A}_I) \subseteq \mathcal{L}(\mathcal{A}_S)$). The *threshold edit distance* (TED) problem is a generalization of the language inclusion problem, which for a given integer threshold $k \geq 0$ asks whether every word in the source language $\mathcal{L}(\mathcal{A}_I)$ has edit distance at most k to the target language $\mathcal{L}(\mathcal{A}_S)$ (with $k = 0$ we have the traditional language inclusion problem). For example, in simulation-based verification of an implementation against a specification, the measured trace may differ slightly from the specification due to inaccuracies in the implementation. Thus, a trace of the implementation may not be in the specification. However, instead of rejecting the implementation, one can quantify the distance between a measured

*This research was funded in part by the European Research Council (ERC) under grant agreement 267989 (QUAREM), by the Austrian Science Fund (FWF) projects S11402-N23 (RiSE) and Z211-N23 (Wittgenstein Award), FWF Grant No P23499- N23, FWF NFN Grant No S11407-N23 (RiSE), ERC Start grant (279307: Graph Games), MSR faculty fellows award, and by the National Science Centre (NCN), Poland under grant 2014/15/D/ST6/04543.

	$\mathcal{C}_2 = \text{DFA}$	$\mathcal{C}_2 = \text{NFA}$	$\mathcal{C}_2 = \text{DPDA}$	$\mathcal{C}_2 = \text{PDA}$
$\mathcal{C}_1 \in \{\text{DFA}, \text{NFA}\}$	PTime	PSpace-c	PTime	
$\mathcal{C}_1 \in \{\text{DPDA}, \text{PDA}\}$		ExpTime-c (Th. 3)	undecidable	

Table 1: Complexity of the language inclusion problem from \mathcal{C}_1 to \mathcal{C}_2 . Our results are boldfaced.

	$\mathcal{C}_2 = \text{DFA}$	$\mathcal{C}_2 = \text{NFA}$	$\mathcal{C}_2 = \text{DPDA}$	$\mathcal{C}_2 = \text{PDA}$
$\mathcal{C}_1 \in \{\text{DFA}, \text{NFA}\}$	coNP-c [3]	PSpace-c [3]	open (Conj. 35)	
$\mathcal{C}_1 \in \{\text{DPDA}, \text{PDA}\}$	coNP-complete (Th. 29)	ExpTime-c (Th. 13)	undecidable (Prop. 31)	

Table 2: Complexity of $\text{FED}(\mathcal{C}_1, \mathcal{C}_2)$. Our results are boldfaced.

	$\mathcal{C}_2 = \text{DFA}$	$\mathcal{C}_2 = \text{NFA}$	$\mathcal{C}_2 = \text{DPDA}$	$\mathcal{C}_2 = \text{PDA}$
$\mathcal{C}_1 \in \{\text{DFA}, \text{NFA}\}$	PSpace-c [2]		undecidable (Prop. 34)	
$\mathcal{C}_1 \in \{\text{DPDA}, \text{PDA}\}$	ExpTime-c (Th. 3 (1))		undecidable	

Table 3: Complexity of $\text{TED}(\mathcal{C}_1, \mathcal{C}_2)$. Our results are boldfaced.

trace and the specification. Among all implementations that violate a specification, the closer the implementation traces are to the specification, the better [6, 8, 13]. The edit distance problem is also the basis for *repairing* specifications [2, 3].

The TED problem answers a fine-grained question with a fixed bound on the number of edit operations. A related problem, the *finite edit distance* (FED) problem, asks whether there exists $k \geq 0$ such that the answer to the TED problem with threshold k is YES. Hence, in verification applications we ask the FED question first, and in case of the positive answer, we can ask the TED question.

Our models. In this work we consider the edit distance computation problem between two automata \mathcal{A}_1 and \mathcal{A}_2 , where \mathcal{A}_1 and \mathcal{A}_2 can be (non-)deterministic finite automata or pushdown automata. Pushdown automata are the standard models for programs with recursion, and regular languages are canonical to express the basic properties of systems that arise in verification. We denote by DPDA (resp., PDA) deterministic (resp., non-deterministic) pushdown automata, and DFA (resp., NFA) deterministic (resp., non-deterministic) finite automata. We consider source and target languages defined by DFA, NFA, DPDA, and PDA. We first present the known results and then our contributions.

Previous results. The main results for the classical language inclusion problem are as follows [14]: (i) if the target language is a DFA, then it can be solved in polynomial time; (ii) if either the target language is a PDA or both source and target languages are DPDA, then it is undecidable; (iii) if the target language is an NFA, then (a) if the source language is a DFA or NFA, then it is **PSpace-complete**, and (b) if the source language is a DPDA or PDA, then it is **PSpace-hard** and can be solved in **ExpTime** (to the best of our knowledge, there is a complexity gap where the upper bound is **ExpTime** and the lower bound is **PSpace**). The TED and FED problems were studied for DFA and NFA. The TED problem is **PSpace-complete**, when the source and target languages are given by DFA or NFA [2, 3]. When the source language is given by a DFA or NFA, the FED problem is: (i) **coNP-complete**, when the target language is given by a DFA [3], (ii) **PSpace-complete**, when the target language is given by an NFA [3].

Our contributions. Our main contributions are as follows.

1. We show that the TED problem is **ExpTime-complete**, when the source language is given by a DPDA or a PDA, and the target language is given by a DFA or NFA. We present a hardness result which shows that the TED problem is **ExpTime-hard** for source languages given as DPDA and target languages given as DFA. We present a matching upper bound by showing that for source languages given as PDA and target languages given as NFA the problem can be solved in **ExpTime**. As a consequence of our lower bound we obtain that the language inclusion problem for source languages given by DPDA (or PDA) and target languages given by NFA is **ExpTime-complete**. In contrast, if the target language is given by a DPDA, then the TED problem is undecidable even for source languages given as DFA. Thus we present a complete picture of the complexity of the TED problem, and in addition we close a complexity gap in the classical language inclusion problem. Note that the interesting verification question is when the implementation (source language) is a DPDA (or PDA) and

the specification (target language) is given as a DFA (or NFA), for which we present decidability results with optimal complexity.

2. We also study the FED problem. For finite automata, it was shown in [2, 3] that if the answer to the FED problem is YES, then a polynomial bound on k exists. In contrast, the edit distance can be exponential between DPDA and DFA. We present a matching exponential upper bound on k for the FED problem from PDA to NFA. We show that when source languages are given as DPDA or PDA, the FED problem is: (i) **coNP**-complete, if the target languages are given as DFA, and (ii) **ExpTime**-complete, if the target languages are given as NFA.

The lower bound in (i) holds even for source languages given as DFA [3]. Our results are summarized in Tables 1, 2 and 3.

This paper extends [7] in the following two ways:

- We provide full proofs of all results from [7].
- We show that the FED problem is **coNP**-complete if the source language is given by DPDA or PDA and the target language is an DFA. This result is technically involved, but it completes the complexity picture for the FED problem in case of the source language given by a pushdown automaton and the target language given by a finite automaton.

Related work. Algorithms for edit distance have been studied extensively for words [16, 1, 19, 20, 15, 18]. The edit distance between regular languages was studied in [2, 3], between timed automata in [9], and between straight line programs in [17, 12]. A near-linear time algorithm to approximate the edit distance for a word to a DYCK language has been presented in [21].

2 Preliminaries

2.1 Words, languages and automata

Words. Given a finite alphabet Σ of letters, a *word* w is a finite sequence of letters. For a word w , we define $w[i]$ as the i -th letter of w and $|w|$ as its length. For instance, if $w = abc$, then $w[2] = b$ and $|w| = 3$. We denote the set of all words over Σ by Σ^* . We use ϵ to denote the empty word.

Pushdown automata. A (*non-deterministic*) *pushdown automaton* (PDA) is a tuple $(\Sigma, \Gamma, Q, S, \delta, F)$, where Σ is the input alphabet, Γ is a finite stack alphabet, Q is a finite set of states, $S \subseteq Q$ is a set of initial states, $\delta \subseteq Q \times \Sigma \times (\Gamma \cup \{\perp\}) \times Q \times \Gamma^*$ is a finite transition relation and $F \subseteq Q$ is a set of final (accepting) states. A PDA $(\Sigma, \Gamma, Q, S, \delta, F)$ is a *deterministic pushdown automaton* (DPDA) if $|S| = 1$ and δ is a function from $Q \times \Sigma \times (\Gamma \cup \{\perp\})$ to $Q \times \Gamma^*$. We denote the class of all PDA (resp., DPDA) by PDA (resp., DPDA). We define the size of a PDA $\mathcal{A} = (\Sigma, \Gamma, Q, S, \delta, F)$, denoted by $|\mathcal{A}|$, as $|Q| + |\delta|$.

Runs of pushdown automata. Given a PDA \mathcal{A} and a word $w = w[1] \dots w[k]$ over Σ , a *run* π of \mathcal{A} on w is a sequence of elements from $Q \times \Gamma^*$ of length $k + 1$ such that $\pi[0] \in S \times \{\epsilon\}$ and for every $i \in \{1, \dots, k\}$ either (1) $\pi[i - 1] = (q, \epsilon)$, $\pi[i] = (q', u')$ and $(q, w[i], \perp, q', u') \in \delta$, or (2) $\pi[i - 1] = (q, ua)$, $\pi[i] = (q', uu')$ and $(q, w[i], a, q', u') \in \delta$. A run π of length $k + 1$ is *accepting* if $\pi[k] \in F \times \{\epsilon\}$, i.e., the automaton is in an accepting state and the stack is empty. The *language recognized (or accepted)* by \mathcal{A} , denoted $\mathcal{L}(\mathcal{A})$, is the set of words that have an accepting run.

Context free grammar (CFG). A context free grammar (CFG) is a tuple (Σ, V, S, P) , where Σ is the alphabet, V is a set of *non-terminals*, $S \in V$ is a *start symbol* and P is a set of *production rules*. A production rule p has the following form $p : A \rightarrow u$, where $A \in V$ and $u \in (\Sigma \cup V)^*$.

A CFG in Chomsky normal form (CNF) is the special case in which each production rule p has one of the following forms (recall that S is the start symbol): (1) $p : A \rightarrow BC$, where $A \in V$ and $B, C \in V \setminus \{S\}$; or (2) $p : A \rightarrow \alpha$, where $A \in V$ and $\alpha \in \Sigma$; or (3) $p : S \rightarrow \epsilon$. It is well-known that any CFG can be brought onto CNF in polynomial time [11].

Languages generated by CFGs. Fix a CFG $G = (\Sigma, V, S, P)$. We define *derivation* \rightarrow_G as a relation on $(\Sigma \cup V)^* \times (\Sigma \cup V)^*$ as follows: $w \rightarrow_G w'$ iff $w = w_1 A w_2$, with $A \in V$, and $w' = w_1 u w_2$ for some $u \in (\Sigma \cup V)^*$ such that $A \rightarrow u$ is a production from G . We define \rightarrow_G^* as the transitive closure of \rightarrow_G . The *language generated* by G , denoted by $\mathcal{L}(G) = \{w \in \Sigma^* \mid S \rightarrow_G^* w\}$ is the set of words that can be derived from S . We omit G and write \rightarrow^* for \rightarrow_G^* if G is clear from the context and for any non-terminal A and word $w \in (\Sigma \cup V)^*$, we call $A \rightarrow^* w$ an

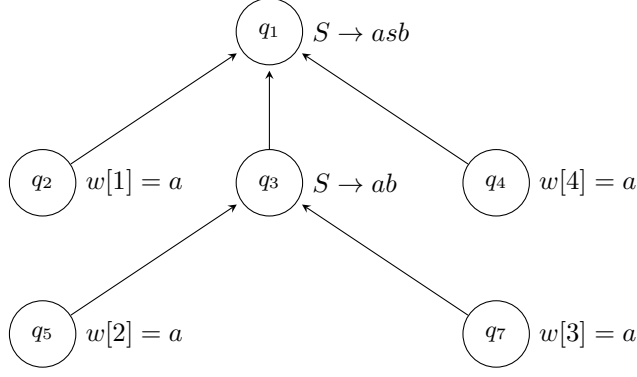


Figure 1: Example of a derivation tree of $w = aabb$ for the CFG G given in paragraph "Languages generated by CFGs."

implied production rule. For instance, the CFG $G = (\Sigma, V, S, P)$, where $\Sigma = \{a, b\}$, $V = \{S\}$, and the rules P are $S \rightarrow aSb$ and $S \rightarrow ab$, generates the language $\{a^n b^n \mid n \geq 1\}$.

It is well-known [14] that CFGs and PDAs are language-wise polynomial equivalent (i.e., there is a polynomial time procedure that, given a PDA, outputs a CFG of the same language and vice versa).

Derivation trees of CFGs. Fix a CFG $G = (\Sigma, V, S, P)$. The CFG defines a (typically infinite) set of *derivation trees*. A derivation tree is an ordered tree¹ where (1) each leaf is associated with an element of $\Sigma \cup V \cup \{\epsilon\}$; and (2) each internal node q is associated with a non-terminal $A \in V$ and production rule $p : A \rightarrow w$, such that A has $|w|$ children and the i -th child, for each i , is associated with $w[i]$ if it is a leaf or a production rule $p' : w[i] \rightarrow w'$ if it is an internal node. A derivation tree T defines a string $w(T)$ over $\Sigma \cup V$ formed by reading labels of the leaves of T in an ascending lexicographic path order ("from left to right") while skipping ϵ symbols. Existence of a derivation tree T with the root A certifies that $A \rightarrow_G^* w(T)$. For instance given G (as in the previous paragraph), the derivation tree for $aabb$ is as given in Figure 1.

Finite automata. A *non-deterministic finite automaton* (NFA) is a pushdown automaton with empty stack alphabet. We will omit Γ while referring to NFA, i.e., we will consider them as tuples $(\Sigma, Q, S, \delta, F)$. We denote the class of all NFA by NFA. Analogously to DPDA we define *deterministic finite automata* (DFA).

Language inclusion. Let $\mathcal{C}_1, \mathcal{C}_2$ be subclasses of PDA. The *inclusion problem from \mathcal{C}_1 in \mathcal{C}_2* asks, given $\mathcal{A}_1 \in \mathcal{C}_1$, $\mathcal{A}_2 \in \mathcal{C}_2$, whether $\mathcal{L}(\mathcal{A}_1) \subseteq \mathcal{L}(\mathcal{A}_2)$.

Single letter operations on words. A single letter operation on a word can be either an *insertion*, a *deletion*, or a *substitution*. Given a letter $a \in \Sigma$ and a number i we define relations $\rightarrow_{I(a,i)}, \rightarrow_{D(a,i)}, \rightarrow_{S(a,i)} \subseteq \Sigma^* \times \Sigma^*$ as follows

- the insert relation $\rightarrow_{I(a,i)}$: for all w, w' we have $w \rightarrow_{I(a,i)} w'$ iff $w' = w[1] \dots w[i]aw[i+1] \dots w[|w|]$. For example, $abc \rightarrow_{I(a,2)} abac$.
- the delete relation $\rightarrow_{D(a,i)}$: for all w, w' we have $w \rightarrow_{D(a,i)} w'$ iff $w' = w[1] \dots w[i-1]w[i+1] \dots w[|w|]$. For example, $abc \rightarrow_{D(b,2)} ac$. (Note that we ignore the letter parameter for deletions. We use $\rightarrow_{D(a,i)}$ over a notation like $\rightarrow_{D(i)}$ to ensure that all three types of single letter operations have 2 parameters)
- the substitution relation $\rightarrow_{S(a,i)}$: for all w, w' we have $w \rightarrow_{S(a,i)} w'$ iff $w' = w[1] \dots w[i-1]aw[i+1] \dots w[|w|]$. For example, $abc \rightarrow_{S(a,2)} aac$.

Edit distance between words. Given two words w_1, w_2 , the edit distance between w_1, w_2 , denoted by $ed(w_1, w_2)$, is the minimal number of single letter operations: insertions, deletions, and substitutions, necessary to transform w_1 into w_2 . More formally, $k := ed(w_1, w_2)$ is the length of the shortest sequence $S_1 S_2 \dots S_k$, where each S_j is an operation $S_j = (P_j, a_j, i_j) \in \{I, D, S\} \times \Sigma \times \mathbb{N}$ for each j , such that there exist words $s_i, i \in \{0, \dots, k\}$, for which (1) $w_1 = s_0$, (2) $w_2 = s_k$ and (3) $s_{j-1} \rightarrow_{P_j(a_j, i_j)} s_j$ for all $j \in \{1, \dots, k\}$.

¹In an ordered tree, children of every node are ordered.

Edit distance between languages. Let $\mathcal{L}_1, \mathcal{L}_2$ be languages. We define the edit distance from \mathcal{L}_1 to \mathcal{L}_2 , denoted $ed(\mathcal{L}_1, \mathcal{L}_2)$, as $\sup_{w_1 \in \mathcal{L}_1} \inf_{w_2 \in \mathcal{L}_2} ed(w_1, w_2)$. The edit distance between languages is not a distance function. In particular, it is not symmetric. For example: $ed(\{a\}^*, \{a, b\}^*) = 0$, while $ed(\{a, b\}^*, \{a\}^*) = \infty$ because for every n , we have $ed(\{b^n\}, \{a\}^*) = n$.

2.2 Problem statement

In this section we define the problems of interest. Then, we recall the previous results and succinctly state our results.

Definition 1. For $\mathcal{C}_1, \mathcal{C}_2 \in \{\text{DFA}, \text{NFA}, \text{DPDA}, \text{PDA}\}$ we define the following questions:

1. The threshold edit distance problem from \mathcal{C}_1 to \mathcal{C}_2 (denoted $\text{TED}(\mathcal{C}_1, \mathcal{C}_2)$): Given automata $\mathcal{A}_1 \in \mathcal{C}_1$, $\mathcal{A}_2 \in \mathcal{C}_2$ and an integer threshold $k \geq 0$, decide whether $ed(\mathcal{L}(\mathcal{A}_1), \mathcal{L}(\mathcal{A}_2)) \leq k$.
2. The finite edit distance problem from \mathcal{C}_1 to \mathcal{C}_2 (denoted $\text{FED}(\mathcal{C}_1, \mathcal{C}_2)$): Given automata $\mathcal{A}_1 \in \mathcal{C}_1$, $\mathcal{A}_2 \in \mathcal{C}_2$, decide whether $ed(\mathcal{L}(\mathcal{A}_1), \mathcal{L}(\mathcal{A}_2)) < \infty$.
3. Computation of edit distance from \mathcal{C}_1 to \mathcal{C}_2 : Given automata $\mathcal{A}_1 \in \mathcal{C}_1$, $\mathcal{A}_2 \in \mathcal{C}_2$, compute $ed(\mathcal{L}(\mathcal{A}_1), \mathcal{L}(\mathcal{A}_2))$.

We establish the complete complexity picture for the TED problem for all combinations of source and target languages given by DFA, NFA, DPDA and PDA:

1. TED for regular languages has been studied in [2], where PSpace-completeness of $\text{TED}(\mathcal{C}_1, \mathcal{C}_2)$ for $\mathcal{C}_1, \mathcal{C}_2 \in \{\text{DFA}, \text{NFA}\}$ has been established.
2. In Section 3, we study the TED problem for source languages given by pushdown automata and target languages given by finite automata. We establish ExpTime-completeness of $\text{TED}(\mathcal{C}_1, \mathcal{C}_2)$ for $\mathcal{C}_1 \in \{\text{DPDA}, \text{PDA}\}$ and $\mathcal{C}_2 \in \{\text{DFA}, \text{NFA}\}$.
3. In Section 5, we study the TED problem for target languages given by pushdown automata. We show that $\text{TED}(\mathcal{C}_1, \mathcal{C}_2)$ is undecidable for $\mathcal{C}_1 \in \{\text{DFA}, \text{NFA}, \text{DPDA}, \text{PDA}\}$ and $\mathcal{C}_2 \in \{\text{DPDA}, \text{PDA}\}$.

We study the FED problem for all combinations of source and target languages given by DFA, NFA, DPDA and PDA and obtain the following results:

1. FED for regular languages has been studied in [3]. It has been shown that for $\mathcal{C}_1 \in \{\text{DFA}, \text{NFA}\}$, the problem $\text{FED}(\mathcal{C}_1, \text{DFA})$ is coNP-complete, while the problem $\text{FED}(\mathcal{C}_1, \text{NFA})$ is PSpace-complete.
2. We show in Section 4 that for $\mathcal{C}_1 \in \{\text{DPDA}, \text{PDA}\}$, the problem $\text{FED}(\mathcal{C}_1, \text{NFA})$ is ExpTime-complete and the problem $\text{FED}(\mathcal{C}_1, \text{DFA})$ is coNP-complete.
3. We show in Section 5 that (1) for $\mathcal{C}_1 \in \{\text{DFA}, \text{NFA}, \text{DPDA}, \text{PDA}\}$, the problem $\text{FED}(\mathcal{C}_1, \text{PDA})$ is undecidable, and (2) the problem $\text{FED}(\text{DPDA}, \text{DPDA})$ is undecidable.

Remark 2. Weighted edit-distance. One could also consider a notion of weighted edit-distance, where a weight function $f : \{I, D, S\} \times \Sigma \rightarrow \mathbb{Z}$ is given that to each edit operation and letter assigns a weight. I.e. inserting a letter a might have a different weight from inserting a letter b . The weighted edit-distance $wed(w_1, w_2)$ would then be the minimum sum of weights $\sum_{j=1}^k f(P_j, a_j)$ over any k and sequence of edit operation $S_1 \dots S_k$, where $S_j = (P_j, a_j, i_j) \in \{I, D, S\} \times \Sigma \times \mathbb{N}$ for each j , such that there exists words s_i , $i \in \{0, \dots, k\}$, for which (1) $w_1 = s_0$, (2) $w_2 = s_k$ and (3) $s_{j-1} \rightarrow_{P_j(a_j, i_j)} s_j$ for all $j \in \{1, \dots, k\}$.

Our results extend to the case where f assigns **positive** weights. There are naturally no differences for the FED case (since if the minimum length is infinite, then so too is the sum of weights). There are no differences either for the TED case, since the only time it comes up (in the following Claim 5) there are no differences.

Allowing f to assign zero or infinite weights leads to distances very different from the classical edit distance, such as the Humming distance, or the length difference. Such distances are out of scope of this paper.

3 Threshold edit distance from pushdown to regular languages

In this section we establish the complexity of the TED problem from pushdown to finite automata.

Theorem 3. (1) For $\mathcal{C}_1 \in \{\text{DPDA}, \text{PDA}\}$ and $\mathcal{C}_2 \in \{\text{DFA}, \text{NFA}\}$, the $\text{TED}(\mathcal{C}_1, \mathcal{C}_2)$ problem is ExpTime-complete. (2) For $\mathcal{C}_1 \in \{\text{DPDA}, \text{PDA}\}$, the language inclusion problem from \mathcal{C}_1 in NFA is ExpTime-complete.

We establish the above theorem as follows: In Section 3.1, we present an exponential-time algorithm for TED(PDA, NFA) (for the upper bound of (1)). Then, in Section 3.2 we show (2), in a slightly stronger form, and reduce it (that stronger problem), to TED(DPDA, DFA), which shows the ExpTime-hardness part of (1). We conclude this section with a brief discussion on parametrized complexity of TED in Section 3.3.

3.1 Upper bound

We present an ExpTime algorithm that, given (1) a PDA \mathcal{A}_P ; (2) an NFA \mathcal{A}_N ; and (3) a threshold t given in binary, decides whether the edit distance from \mathcal{A}_P to \mathcal{A}_N is above t . The algorithm extends a construction for NFA by Benedikt et al. [2].

Intuition. The construction uses the idea that for a given word w and an NFA \mathcal{A}_N the following are equivalent: (i) $ed(w, \mathcal{A}_N) > t$, and (ii) for each accepting state s of \mathcal{A}_N and for every word w' , if \mathcal{A}_N can reach s from some initial state upon reading w' , then $ed(w, w') > t$. We construct a PDA \mathcal{A}_I which simulates the PDA \mathcal{A}_P and stores in its states all states of the NFA \mathcal{A}_N reachable with at most t edits. More precisely, the PDA \mathcal{A}_I remembers in its states, for every state s of the NFA \mathcal{A}_N , the minimal number of edit operations necessary to transform the currently read prefix w_p of the input word into a word w'_p , upon which \mathcal{A}_N can reach s from some initial state. If for some state the number of edit operations exceeds t , then we associate with this state a special symbol $\#$ to denote this. Then, we show that a word w accepted by the PDA \mathcal{A}_P has $ed(w, \mathcal{A}_N) > t$ iff the automaton \mathcal{A}_I has a run on w that ends (1) in an accepting state of simulated \mathcal{A}_P , (2) with the simulated stack of \mathcal{A}_P empty, and (3) the symbol $\#$ is associated with every accepting state of \mathcal{A}_N .

Lemma 4. *Given (1) a PDA \mathcal{A}_P ; (2) an NFA \mathcal{A}_N ; and (3) a threshold t given in binary, the decision problem of whether $ed(\mathcal{A}_P, \mathcal{A}_N) \leq t$ can be reduced to the emptiness problem for a PDA of size $O(|\mathcal{A}_P| \cdot (t + 2)^{|\mathcal{A}_N|})$.*

Proof. Let Q_N (resp., F_N) be the set of states (resp., accepting states) of \mathcal{A}_N . For $i \in \mathbb{N}$ and a word w , we define $T_w^i = \{s \in Q_N : \text{there exists } w' \text{ with } ed(w, w') = i \text{ such that } \mathcal{A}_N \text{ has a run on the word } w' \text{ ending in } s\}$. For a pair of states $s, s' \in Q_N$ and $\alpha \in \Sigma \cup \{\epsilon\}$, we define $m(s, s', \alpha)$ as the minimum number of edits needed to apply to α so that \mathcal{A}_N has a run on the resulting word from s' to s . For all $s, s' \in Q_N$ and $\alpha \in \Sigma \cup \{\epsilon\}$, we can compute $m(s, s', \alpha)$ in polynomial time in $|\mathcal{A}_N|$. For a state $s \in Q_N$ and a word w let $d_w^s = \min\{i \geq 0 \mid s \in T_w^i\}$, i.e., d_w^s is the minimal number of edits necessary to apply to w such that \mathcal{A}_N reaches s upon reading the resulting word. We will first prove the following claim.

Claim 5. *We have that $d_{wa}^s = \min_{s' \in Q_N} (d_w^{s'} + m(s, s', a))$*

Proof. Consider a run witnessing d_{wa}^s . As shown by [22] we can split the run into two parts, one sub-run on w ending in s' , for some s' , and one sub-run on a starting in s' . Clearly, the sub-run on w has used $d_w^{s'}$ edits and the one on a has used $m(s, s', a)$ edits. \square

Let Q_P (resp., F_P) be the set of states (resp., accepting states) of the PDA \mathcal{A}_P . For every word w and every state $q \in Q_P$ such that there is a run on w ending in q , we define $\text{Impact}(w, q, \mathcal{A}_P, \mathcal{A}_N, t)$ as a pair (q, λ) in $Q_P \times \{0, 1, \dots, t, \#\}^{|Q_N|}$, where λ is defined as follows: for every $s \in Q_N$ we have $\lambda(s) = d_w^s$ if $d_w^s \leq t$, and $\lambda(s) = \#$ otherwise. Clearly, the edit distance from \mathcal{A}_P to \mathcal{A}_N exceeds t if there is a word w and an accepting state q of \mathcal{A}_P such that $\text{Impact}(w, q, \mathcal{A}_P, \mathcal{A}_N, t)$ is a pair (q, λ) and for every $s \in F_N$ we have $\lambda(s) = \#$ (i.e., the word w is in $\mathcal{L}(\mathcal{A}_P)$ but any run of \mathcal{A}_N ending in F_N has distance exceeding t).

We can now construct an *impact automaton*, a PDA \mathcal{A}_I , with state space $Q_P \times \{0, 1, \dots, t, \#\}^{|Q_N|}$ and the transition relation defined as follows: A tuple $(\langle q, \lambda_1 \rangle, a, \gamma, \langle q', \lambda_2 \rangle, u)$ is a transition of \mathcal{A}_I iff the following conditions hold:

1. the tuple projected to the first component of its state (i.e., the tuple (q, a, γ, q', u)) is a transition of \mathcal{A}_P , and
2. the second component λ_2 is computed from λ_1 using Claim 5, i.e., for every $s \in Q_N$ we have $\lambda_2(s) = \min_{s' \in Q_N} (\lambda_1(s') + m(s, s', a))$.

The initial states of \mathcal{A}_I are $S_P \times \{\lambda_0\}$, where S_P are initial states of \mathcal{A}_P and λ_0 is defined as follows. For every $s \in Q_N$ we have $\lambda_0(s) = \min_{s' \in S_N} m(s, s', \epsilon)$, where S_N are initial states of \mathcal{A}_N (i.e., a start state of \mathcal{A}_I is a pair of a start state of \mathcal{A}_P together with the vector where the entry describing s is the minimum number of edits needed

to get to the state s on the empty word). Also, the accepting states are $\{\langle q, \lambda \rangle \mid q \in F_P \text{ and for every } s \in F_N \text{ we have } \lambda(s) = \#\}$. Observe that for a run of \mathcal{A}_I on w ending in (s, λ) , the vector $\text{Impact}(w, s, \mathcal{A}_P, \mathcal{A}_N, t)$ is precisely (s, λ) . Thus, the PDA \mathcal{A}_I accepts a word w iff the edit distance between \mathcal{A}_P and \mathcal{A}_N is above t . Since the size of \mathcal{A}_I is $O(|\mathcal{A}_P| \cdot (t+2)^{|\mathcal{A}_N|})$ we obtain the desired result. \square

Lemma 4 implies the following:

Lemma 6. *TED(PDA, NFA) is in ExpTime.*

Proof. Let $\mathcal{A}_P, \mathcal{A}_N$ and t be an instance of TED(PDA, NFA), where \mathcal{A}_P is a PDA, \mathcal{A}_N is an NFA, and t is a threshold given in binary. By Lemma 4, we can reduce TED to the emptiness question of a PDA of the size $O(|\mathcal{A}_P| \cdot (t+2)^{|\mathcal{A}_N|})$. Since $|\mathcal{A}_P| \cdot (t+2)^{|\mathcal{A}_N|}$ is exponential in $|\mathcal{A}_P| + |\mathcal{A}_N| + t$ and the emptiness problem for PDA can be decided in time polynomial in their size [14], the result follows. \square

3.2 Lower bound

Our ExpTime-hardness proof of TED(DPDA, DFA) extends the idea from [2] that shows PSpace-hardness of the edit distance for DFA. The standard proof of PSpace-hardness of the universality problem for NFA [14] is by reduction to the halting problem of a fixed Turing machine M working on a bounded tape. The Turing machine M is the one that simulates other Turing machines (such a machine is called universal). The input to that problem is the initial configuration C_1 and the tape is bounded by its size $|C_1|$. In the reduction, the NFA recognizes the language of all words that do not encode a valid computation of M starting from the initial configuration C_1 , i.e., it accepts if one of the following conditions is violated: (1) the given word is a sequence of configurations, (2) the state of the Turing machine and the adjacent letters follow from transitions of M , (3) the first configuration is C_1 and (4) the tape's cells are changed only by M , i.e., they do not change values spontaneously. While violation of conditions (1), (2) and (3) can be checked by a DFA of polynomial size, condition (4) can be encoded by a polynomial-size NFA but not a polynomial-size DFA. However, to check (4) the automaton has to make only a single non-deterministic choice to pick a position in the encoding of the computation, which violates (4), i.e., the value at that position is different from the value $|C_1| + 1$ letters further, which corresponds to the same memory cell in the successive configuration, and the head of M does not change it. We can transform a non-deterministic automaton \mathcal{A}_N checking (4) into a deterministic automaton \mathcal{A}_D by encoding such a non-deterministic pick using an external letter. Since we need only one external symbol, we show that $\mathcal{L}(\mathcal{A}_N) = \Sigma^*$ iff $\text{ed}(\Sigma^*, \mathcal{L}(\mathcal{A}_D)) = 1$. This suggests the following definition:

Definition 7. *An NFA $\mathcal{A} = (\Sigma, Q, S, \delta, F)$ is nearly-deterministic if $|S| = 1$ and $\delta = \delta_1 \cup \delta_2$, where δ_1 is a function and in every accepting run the automaton takes a transition from δ_2 exactly once.*

Lemma 8. *There exists a DPDA \mathcal{A}_P such that the problem, given a nearly-deterministic NFA \mathcal{A}_N , decide whether $\mathcal{L}(\mathcal{A}_P) \subseteq \mathcal{L}(\mathcal{A}_N)$, is ExpTime-hard.*

Proof. Consider the linear-space halting problem for a (fixed) alternating Turing machine (ATM) M : given an input word w over an alphabet Σ , decide whether M halts on w with the tape bounded by $|w|$. There exists an ATM M_U , such that the linear-space halting problem for M_U is ExpTime-complete [5]. We show the ExpTime-hardness of the problem from the lemma statement by reduction from the linear-space halting problem for M_U .

Without loss of generality, we assume that existential and universal transitions of M_U alternate. Fix an input of length n . The main idea is to construct the language L of words that encode valid terminating computation trees of M_U on the given input. Observe that the language L depends on the given input. We encode a single configuration of M_U as a word of length $n+1$ of the form $\Sigma^i q \Sigma^{n-i}$, where q is a state of M_U . Recall that a computation of an ATM is a tree, where every node of the tree is a configuration of M_U , and it is accepting if every leaf node is an accepting configuration. We encode computation trees T of M_U by traversing T in pre-order and executing the following: if the current node has only one successor, then write down the current configuration C , terminate it with $\#$ and move down to the successor node in T . Otherwise, if the current node has two successors s, t in the tree, then write down in order (1) the reversed current configuration C^R ; and (2) the results of traversals on s and t , each surrounded by parentheses

(and), i.e., $C_1^R (u^s) (u^t)$, where u^s (resp., u^t) is the result of the traversal of the sub-tree of T rooted at s (resp., t). Finally, if the current node is a leaf, write down the corresponding configuration and terminate with $\$$. For example, consider a computation with the initial configuration C_1 , from which an existential transition leads to C_2 , which in turn has a universal transition to C_3 and C_4 . Such a computation tree is encoded as follows:

$$C_1 \# C_2^R (C_3 \dots \$) (C_4 \dots \$).$$

We define automata \mathcal{A}_N and \mathcal{A}_P over the alphabet $\Sigma \cup \{\#, \$, (,)\}$. The automaton \mathcal{A}_N is a nearly deterministic NFA that recognizes only (but not all) words not encoding valid computation trees of M_U . More precisely, \mathcal{A}_N accepts in four cases: (1) The word does not encode a tree (except that the parentheses may not match as the automaton cannot check that) of computation as presented above. (2) The initial configuration is different from the one given as the input. (3) The successive configurations, i.e., those that result from existential transitions or left-branch universal transitions (like C_2 to C_3), are not valid. The right-branch universal transitions, which are preceded by the word “(”, are not checked by \mathcal{A}_N . For example, the consistency of the transition C_2 to C_4 is not checked by \mathcal{A}_N . Finally, (4) \mathcal{A}_N accepts words in which at least one final configuration, which is a configuration followed by $\$$, is not final for M_U . Observe that conditions (1), (2) and (4) can be checked by polynomial-size DFA. Condition (3) can be checked by a polynomial-size nearly-deterministic NFA, which picks a position in C_2 , for which the corresponding position in C_3 is faulty (either contains a spontaneous change of the corresponding tape cell or it is not compatible with any transition of M_U). Picking such a position correspond to taking transition δ_2 by a nearly-deterministic NFA. Thus, the automaton \mathcal{A}_N is a nearly deterministic NFA, which recognizes the union of automata recognizing (1)-(4).

Next, we define \mathcal{A}_P as a DPDA that accepts words in which parentheses match and right-branch universal transitions are consistent, e.g., it checks consistency of a transition from C_2 to C_4 . The automaton \mathcal{A}_P pushes configurations on even levels of the computation tree (e.g., C_2^R), which are reversed, on the stack and pops these configurations from the stack to compare them with the following configuration in the right sub-tree (e.g., C_4). In the example this means that, while the automaton processes the sub-word $(C_3 \dots \$)$, it can use its stack to check consistency of universal transitions in that sub-word. We assumed that M_U does not have consecutive universal transitions. This means that, for example, \mathcal{A}_P does not need to check the consistency of C_4 with its successive configuration. By construction, we have $L = \mathcal{L}(\mathcal{A}_P) \cap \mathcal{L}(\mathcal{A}_N)^c$ (recall that L is the language of encodings of computations of M_U on the given input) and M_U halts on the given input if and only if $\mathcal{L}(\mathcal{A}_P) \subseteq \mathcal{L}(\mathcal{A}_N)$ fails. Observe that \mathcal{A}_P is fixed for all inputs, since it only depends on the fixed Turing machine M_U . \square

Now, the following lemma, which is (2) of Theorem 3, follows from Lemma 8.

Lemma 9. *The language inclusion problem from DPDA to NFA is ExpTime-complete.*

Proof. The ExpTime upper bound is immediate (basically, an exponential determinization of the NFA, followed by complementation, product construction with the PDA, and the emptiness check of the product PDA in polynomial time in the size of the product). ExpTime-hardness of the problem follows from Lemma 8. \square

Now, we show that the inclusion problem of DPDA in nearly-deterministic NFA, which is ExpTime-complete by Lemma 8, reduces to TED(DPDA, DFA). In the reduction, we transform a nearly-deterministic NFA \mathcal{A}_N over the alphabet Σ into a DFA \mathcal{A}_D by encoding a single non-deterministic choice by auxiliary letters.

Lemma 10. *TED(DPDA, DFA) is ExpTime-hard.*

Proof. To show ExpTime-hardness of TED(DPDA, DFA), we reduce the inclusion problem of DPDA in nearly-deterministic NFA to TED(DPDA, DFA). Consider a DPDA \mathcal{A}_P and a nearly-deterministic NFA \mathcal{A}_N over an alphabet Σ . Without loss of generality we assume that letters on even positions are $\diamond \in \Sigma$ and \diamond do not appear on the odd positions. Let $\delta = \delta_1 \cup \delta_2$ be the transition relation of \mathcal{A}_N , where δ_1 is a function and along each accepting run, \mathcal{A}_N takes exactly one transition from δ_2 . We transform the NFA \mathcal{A}_N to a DFA \mathcal{A}_D by extending the alphabet Σ with external letters $\{1, \dots, |\delta_2|\}$. On letters from Σ , the automaton \mathcal{A}_D takes transitions from δ_1 . On a letter $i \in \{1, \dots, |\delta_2|\}$, the automaton \mathcal{A}_D takes the i -th transition from δ_2 .

We claim that $\mathcal{L}(\mathcal{A}_P) \subseteq \mathcal{L}(\mathcal{A}_N)$ iff $ed(\mathcal{L}(\mathcal{A}_P), \mathcal{L}(\mathcal{A}_D)) = 1$. Every word $w \in \mathcal{L}(\mathcal{A}_D)$ contains a letter $i \in \{1, \dots, |\delta_2|\}$, which does not belong to Σ . Therefore, $ed(\mathcal{L}(\mathcal{A}_P), \mathcal{L}(\mathcal{A}_D)) \geq 1$. But, if we substitute letter i by the

letter in the i -th transition of δ_2 , we get a word from $\mathcal{L}(\mathcal{A}_N)$. If we simply delete the letter i , we get a word which does not belong to $\mathcal{L}(\mathcal{A}_N)$ as it has letter \diamond on an odd position. Therefore, $ed(\mathcal{L}(\mathcal{A}_P), \mathcal{L}(\mathcal{A}_D)) \leq 1$ implies $\mathcal{L}(\mathcal{A}_P) \subseteq \mathcal{L}(\mathcal{A}_N)$. Finally, consider a word $w' \in \mathcal{L}(\mathcal{A}_N)$. The automaton \mathcal{A}_N has an accepting run on w' , which takes exactly once a transition from δ_2 . Say the taken transition is the i -th transition and the position in w' is p . Then, the word w , obtained from w' by substituting the letter at position p by letter i , is accepted by \mathcal{A}_D . Therefore, $\mathcal{L}(\mathcal{A}_P) \subseteq \mathcal{L}(\mathcal{A}_N)$ implies $ed(\mathcal{L}(\mathcal{A}_P), \mathcal{L}(\mathcal{A}_D)) \leq 1$. Thus we have $\mathcal{L}(\mathcal{A}_P) \subseteq \mathcal{L}(\mathcal{A}_N)$ iff $ed(\mathcal{L}(\mathcal{A}_P), \mathcal{L}(\mathcal{A}_D)) = 1$. \square

3.3 Parameterized complexity

Problems of high complexity can be practically viable if the complexity is caused by a parameter, which tends to be small in the applications. In this section we discuss the dependence of the complexity of TED based on its input values.

Proposition 11. (1) *There exist a threshold $t > 0$ and a DPDA \mathcal{A}_P such that the variant of TED(DPDA, DFA), in which the threshold is fixed to t and DPDA is fixed to \mathcal{A}_P , is still ExpTime-complete.* (2) *The variant of TED(PDA, NFA), in which the threshold is given in unary and NFA is fixed, is in PTime.*

Proof. (1): The inclusion problem of DPDA in nearly-deterministic NFA is ExpTime-complete even if a DPDA is fixed (Lemma 8). Therefore, the reduction in Lemma 10 works for threshold 1 and fixed DPDA.

(2): In the reduction from Lemma 4, the resulting PDA has size $|\mathcal{A}_P| \cdot (t + 2)^{|\mathcal{A}_N|}$, where \mathcal{A}_P is a PDA, \mathcal{A}_N is an NFA and t is a threshold. If \mathcal{A}_N is fixed and t is given in unary, then $|\mathcal{A}_P| \cdot (t + 2)^{|\mathcal{A}_N|}$ is polynomial in the size of the input and we can decide its non-emptiness in polynomial time. \square

Conjecture 12 completes the study of the parametrized complexity of TED.

Conjecture 12. *The variant of TED(PDA, NFA), in which the threshold is given in binary and NFA is fixed, is in PTime.*

4 Finite edit distance from pushdown to regular languages

In this section we study the complexity of the FED problem from pushdown automata to finite automata.

Theorem 13. (1) *For $\mathcal{C}_1 \in \{\text{DPDA}, \text{PDA}\}$ and $\mathcal{C}_2 \in \{\text{DFA}, \text{NFA}\}$ we have the following dichotomy: for all $\mathcal{A}_1 \in \mathcal{C}_1, \mathcal{A}_2 \in \mathcal{C}_2$ either $ed(\mathcal{L}(\mathcal{A}_1), \mathcal{L}(\mathcal{A}_2))$ is exponentially bounded in $|\mathcal{A}_1| + |\mathcal{A}_2|$ or $ed(\mathcal{L}(\mathcal{A}_1), \mathcal{L}(\mathcal{A}_2))$ is infinite. Conversely, for every n there exist a DPDA \mathcal{A}_P and a DFA \mathcal{A}_D , both of the size $O(n)$, such that $ed(\mathcal{L}(\mathcal{A}_P), \mathcal{L}(\mathcal{A}_D))$ is finite and exponential in n (i.e., the dichotomy is asymptotically tight).* (2) *For $\mathcal{C}_1 \in \{\text{DPDA}, \text{PDA}\}$ the FED(\mathcal{C}_1 , NFA) problem is ExpTime-complete.* (3) *For $\mathcal{C}_1 \in \{\text{DPDA}, \text{PDA}\}$ the FED(\mathcal{C}_1 , DFA) problem is coNP-complete.* (4) *Given a PDA \mathcal{A}_P and an NFA \mathcal{A}_N , we can compute the edit distance $ed(\mathcal{L}(\mathcal{A}_P), \mathcal{L}(\mathcal{A}_N))$ in time exponential in $|\mathcal{A}_P| + |\mathcal{A}_N|$.*

First, we show in Section 4.1 the dichotomy of (1), which together with Theorem 3, implies the ExpTime upper bound for (2). Next, in Section 4.2, we show that FED(PDA, DFA) problem is in coNP, which together with the results from [3] shows (3). Finally, in Section 4.3, we show that FED(DPDA, NFA) is ExpTime-hard. We also present the exponential lower bound for (1). Conditions (1), (2), and Theorem 3 imply (3) (by iteratively testing with increasing thresholds upto exponential bounds along with the decision procedure from Theorem 3).

4.1 Upper bound for NFA

In this section we consider the problem of deciding whether the edit distance from a PDA to an NFA is finite.

We first give an overview of the section. Let \mathcal{A}_N be an NFA and \mathcal{A}_P a PDA that has T non-terminals. We show (in Lemma 18) that for any word $w \in \mathcal{L}(\mathcal{A}_P)$ one can break the word into chunks $w = s_1 u_1 \dots s_k u_k s_{k+1}$, such that $\sum_{i=1}^k |s_i| \leq 2^T$ and for any ℓ word w_ℓ defined as $w_\ell = s_1(u_1^\ell) \dots s_k(u_k^\ell) s_{k+1}$ belongs to $\mathcal{L}(\mathcal{A}_P)$ (this is in some sense the opposite of the pumping lemma, since the part that *cannot* be pumped is small). We then show

(this follows from Lemma 19) that if there is a word $w \in \mathcal{L}(\mathcal{A}_P)$ such that $ed(w, \mathcal{L}(\mathcal{A}_N)) > 2^T$, then for every word w_ℓ defined as above we have $ed(w_{\ell+1}, \mathcal{L}(\mathcal{A}_N)) > ed(w_\ell, \mathcal{L}(\mathcal{A}_N))$ for all $\ell \geq 0$, showing that the edit-distance $ed(\mathcal{L}(\mathcal{A}_P), \mathcal{L}(\mathcal{A}_N))$ is unbounded. On the other hand, clearly, if $ed(w, \mathcal{L}(\mathcal{A}_N)) \leq 2^T$ for all $w \in \mathcal{L}(\mathcal{A}_P)$, then the edit-distance $ed(\mathcal{L}(\mathcal{A}_P), \mathcal{L}(\mathcal{A}_N)) \leq 2^T$ by definition.

We start with a reduction of the problem. Given a language \mathcal{L} , we define $\text{prefix}(\mathcal{L}) = \{u : u \text{ is a prefix of some word from } \mathcal{L}\}$. We call an automaton \mathcal{A} a *safety automaton* if every state of \mathcal{A} is accepting. Note that automata are not necessarily total, i.e. some states might not have an outgoing transition for some input symbols, and thus a safety automaton does not necessarily accept all words. Note that for every NFA \mathcal{A}_N , the language $\text{prefix}(\mathcal{L}(\mathcal{A}_N))$ is the language of a safety NFA. We show that $\text{FED}(\text{PDA}, \text{NFA})$ reduces to FED from PDA to safety NFA.

Lemma 14. *Let \mathcal{A}_P be a PDA and \mathcal{A}_N an NFA. The following inequalities hold:*

$$ed(\mathcal{L}(\mathcal{A}_P), \mathcal{L}(\mathcal{A}_N)) \geq ed(\mathcal{L}(\mathcal{A}_P), \text{prefix}(\mathcal{L}(\mathcal{A}_N))) \geq ed(\mathcal{L}(\mathcal{A}_P), \mathcal{L}(\mathcal{A}_N)) - |\mathcal{A}_N|$$

Proof. Since $\mathcal{L}(\mathcal{A}_N) \subseteq \text{prefix}(\mathcal{L}(\mathcal{A}_N))$, we have

$$ed(\mathcal{L}(\mathcal{A}_P), \mathcal{L}(\mathcal{A}_N)) \geq ed(\mathcal{L}(\mathcal{A}_P), \text{prefix}(\mathcal{L}(\mathcal{A}_N)))$$

as the latter is the minimum over a larger set by definition.

Hence, we only need to show the other inequality. First observe that for every $w \in \text{prefix}(\mathcal{L}(\mathcal{A}_N))$, upon reading w , the automaton \mathcal{A}_N can reach a state from which an accepting state is reachable and thus, an accepting state can be reached in at most $|\mathcal{A}_N|$ steps. Therefore, for every $w \in \text{prefix}(\mathcal{L}(\mathcal{A}_N))$ there exists w' of length bounded by $|\mathcal{A}_N|$ such that $ww' \in \mathcal{L}(\mathcal{A}_N)$. It follows that $ed(\mathcal{L}(\mathcal{A}_P), \text{prefix}(\mathcal{L}(\mathcal{A}_N))) \geq ed(\mathcal{L}(\mathcal{A}_P), \mathcal{L}(\mathcal{A}_N)) - |\mathcal{A}_N|$. \square

Remark 15. *Consider an NFA \mathcal{A}_N recognizing a language such that $\text{prefix}(\mathcal{L}(\mathcal{A}_N)) = \Sigma^*$. For every PDA \mathcal{A}_P , the edit distance $ed(\mathcal{L}(\mathcal{A}_P), \mathcal{L}(\mathcal{A}_N))$ is bounded by $|\mathcal{A}_N|$.*

In the remainder of this section we work with context-free grammars (CFGs) instead of PDAs. There are polynomial-time transformations between CFGs and PDAs that preserve the generated language; switching from PDAs to CFGs is made only to simplify the proofs. The following definition and lemma can be seen as a reverse version of the pumping lemma for context free grammars (in that we ensure that the part which can not be pumped is small).

As an abuse of notation we will think of a sequence of words as both the concatenation of the words and the sequence. We define $[1, k] = \{1, \dots, k\}$.

Left and right language. For a CFG G and a non-terminal A , we define the languages

$$\mathcal{L}(G, A, L) = \{w \in \Sigma^* \mid \exists w' \in \Sigma^* (A \rightarrow^* wAw')\} \quad \text{and} \quad \mathcal{L}(G, A, R) = \{w \in \Sigma^* \mid \exists w' \in \Sigma^* (A \rightarrow^* w'Aw)\}.$$

Also, the set of directions D is $D = \{L, R\}$. We next argue that we can construct a CFG for $\mathcal{L}(G, A, D)$.

Lemma 16. *Given a CFG G , a non-terminal A and a direction D , we can construct in polynomial time a CFG G' for which $\mathcal{L}(G') = \mathcal{L}(G, A, D)$.*

Proof. We describe the construction of a CFG for $\mathcal{L}(G, A, L)$ and the construction for $\mathcal{L}(G, A, R)$ is similar.

To simplify, we consider G to be on CNF. We construct G' as follows: The CFG G' consists of two versions of each non-terminal in G , one with a star and one without. I.e. for each non-terminal $X \in G$, we have the non-terminals X and X^* in G' . The idea is that X^* derives prefixes of words derivable from X in G , which ends just before a A . The productions are then as follows:

- **Non-starred.** Each production of G is also in G' , which defines the productions for the non-starred non-terminals.
- **Starred.** For each production $X \rightarrow BC$ in G , there are productions $X^* \rightarrow BC^*$ and $X^* \rightarrow B^*$ in G' .
- **Additional for A^* .** The non-terminal A^* has the production $A^* \rightarrow \epsilon$ in G' (no other starred non-terminal can produce any terminal).

The start symbol of G' is A^* . We next argue that $\mathcal{L}(G') = \mathcal{L}(G, A, L)$.

$\mathcal{L}(G') \subseteq \mathcal{L}(G, A, L)$. It is easy to see from the productions of G' that the only way to remove a starred non-terminal is to eventually replace a A^* by ϵ . By construction this is the last non-terminal in some prefix w of a word in G with start state A and thus $w \in \mathcal{L}(G, A, L)$.

$\mathcal{L}(G, A, L) \subseteq \mathcal{L}(G')$. Given a word w in $\mathcal{L}(G, A, L)$ by definition there is an implied production rule $A \rightarrow^* wAw'$ (in G) for some w' . Given a derivation tree \mathcal{D} for the implied production rule $A \rightarrow^* wAw'$, it is easy to construct a derivation tree \mathcal{D}' for w in $\mathcal{L}(G')$, indicating that w is in $\mathcal{L}(G')$. The two trees \mathcal{D} and \mathcal{D}' are identical except as follows: For a node v in \mathcal{D}' let $\mathcal{D}(v)$ be the corresponding node in \mathcal{D} . Let ℓ be the leaf in \mathcal{D}' such that $\mathcal{D}(v)$ is the leaf with label A in \mathcal{D} . The production rule of ℓ is $A^* \rightarrow \epsilon$. Then, consider the path π from ℓ to the root of \mathcal{D}' . For each internal node v in π where $\mathcal{D}(v)$ has production rule $X \rightarrow BC$, we have the following:

- **π comes from the left child.** If π goes through the left child, the production rule of v is $X^* \rightarrow B^*$ and the sub-tree under the right child of $\mathcal{D}(v)$ is cut out of \mathcal{D}' (including that v has no right child in this case).
- **π comes from the right child.** If π goes through the right child the production rule of v is $X^* \rightarrow BC^*$.

Then, tree \mathcal{D}' spells the word w and is a derivation tree in G' . Thus $w \in \mathcal{L}(G')$ and the lemma follows. \square

Realizable. Given a CFG G in Chomsky normal form, we define the *realizable CFG* \bar{G} (for clarity we do not define it in CNF) that

- for each production of the form $P : A \rightarrow a$ in G have the production $P : A \rightarrow \epsilon$,
- for each production of the form $P : A \rightarrow BC$ in G have the production $P : A \rightarrow a_L^A B C a_R^A$

and no other productions (the language is then especially over the terminals $\{a_D^A \mid D \in \{L, R\} \wedge A \text{ is a non-terminal}\}$). A sequence is *realizable* if it is a sub-sequence of a word in $\mathcal{L}(\bar{G})$, i.e., it results from deletion of letters from some word of $\mathcal{L}(\bar{G})$.

Lemma 17. Let $a_{D_1}^{A_1} \dots a_{D_k}^{A_k}$ be a realizable sequence in G . Then for every sequence of words $w_1 \in \mathcal{L}(G, A_1, D_1), \dots, w_k \in \mathcal{L}(G, A_k, D_k)$ there exist words s_1, \dots, s_k, s_{k+1} such that $s_1 w_1 \dots s_k w_k s_{k+1}$ belongs to $\mathcal{L}(G)$.

Proof. Let $\alpha = a_{D_1}^{A_1} \dots a_{D_k}^{A_k}$. We consider two cases: Either $\alpha \in \mathcal{L}(\bar{G})$ or not.

The case where $\alpha \in \mathcal{L}(\bar{G})$. Consider a derivation tree \mathcal{D} for α . We translate it into a derivation tree in G for $s_1 w_1 \dots s_k w_k s_{k+1}$, by replacing each production (which are in \bar{G}) of the nodes in \mathcal{D} with (generalized) productions in G .

Each leaf node v corresponds to a production $P : A \rightarrow \epsilon$. By definition there exists a production $P : A \rightarrow a$ in G and we then simply replace P in \bar{G} with P in G .

Each non-leaf node v , with children b and c respectively, corresponds to the use of a production $P : A \rightarrow a_L^A B C a_R^A$, where the a_L^A is the i -th letter and a_R^A the j -th of w for some i, j . By definition of $\mathcal{L}(G, A, D)$ we have that there is a production $P' : A \rightarrow^* w_i w_j' B C w_j' w_i'$ in G for some w_i', w_j' . In this case we replace P in \bar{G} with P' in G . (The word s_{i+1} are concatenation of words w_j' and letters derived by productions $P : A \rightarrow a$ corresponding to $P : A \rightarrow \epsilon$.)

The case where $\alpha \notin \mathcal{L}(\bar{G})$. Find a word $\alpha' \in \mathcal{L}(\bar{G})$ such that α is a sub-sequence of α' (letting p be the sequence of positions defining α from α') and do as above with α' and the sequence of words s' which is an extension of the sequence s of length $|\alpha'|$ by inserting ϵ at the remaining positions (i.e., the extension is such that s is the sub-sequence of s' defined by p). \square

Compact G -decomposition. Given a CFG G with a set of non-terminals of size T and a word $w \in \mathcal{L}(G)$, we define a *compact G -decomposition* of w as $w = s_1 u_1 \dots s_k u_k s_{k+1}$ such that

1. for each u_i , there is an associated terminal $a_{D_i}^{A_i}$, such that the sequence $a_{D_1}^{A_1} \dots a_{D_k}^{A_k}$ is realizable and $u_i \in \mathcal{L}(G, A_i, D_i)$.
2. for all $\ell \in \mathbb{N}$, the word $w_\ell := s_1 (u_1)^\ell s_2 \dots s_k (u_k)^\ell s_{k+1}$ is in $\mathcal{L}(G)$.

$$3. |w_0| = \sum_{i=1}^{k+1} |s_i| \leq 2^T \text{ and } k \leq 2^{T+1} - 2.$$

Lemma 18. *For every CFG G in CNF, every word $w \in \mathcal{L}(G)$ admits a compact G -decomposition.*

Intuition. The proof follows by repeated applications of the principle behind the pumping lemma, until the part which is not pumped is small.

Proof. Fix some ℓ and consider some word w in $\mathcal{L}(G)$ and some derivation tree \mathcal{D} for w . We will greedily construct a compact G -representation, using that we do not give bounds on $|u_i|$.

Greedy traversal and the first two properties. The idea is to consider nodes of \mathcal{D} in a depth first pre-order traversal (ensuring that when we consider some node we have already considered its ancestors). When we consider some node v , we continue with the traversal, unless there exists a descendant u of v , such that $\mathcal{D}[v] = \mathcal{D}[u]$. If there exists such a descendant, let u' be the bottom-most descendant (pick an arbitrary one if there are more than one such bottom-most descendants) such that $A := \mathcal{D}[v] = \mathcal{D}[u']$. We say that (v, u') forms a *pump pair* of w . Consider subword $\alpha_v, \alpha_{u'}$ of w derived by subtrees of \mathcal{D} with roots at v and u' respectively. We can then write α_v as $s\alpha_{u'}s'$ (and hence $A \rightarrow_G^* sAs'$), for some s and s' in the obvious way and s and s' will correspond to u_i and u_j respectively for some $i < j$ (i and j are defined by the traversal that we have already assigned $i - 1$ u 's then we first visit v and then assign s as the u_i and then we return to the parent of v , we have assigned $j - 1$ u 's and assign s' to be u_j).

Furthermore, u_i is associated with a_L^A and u_j is associated with a_R^A . Observe that $A \rightarrow^* u_i A u_j$ implies that $u_i \in \mathcal{L}(G, A, L)$ and $u_j \in \mathcal{L}(G, A, R)$ and we therefore have ensured the first property of compact G -representation. This also shows that we can replace u_i with $(u_i)^\ell$ and u_j with $(u_j)^\ell$ (because, clearly $A \rightarrow^* (u_i)^\ell A (u_j)^\ell$) and the new word is in $\mathcal{L}(G)$. Hence, w_ℓ is in $\mathcal{L}(G)$, showing the second property of compact G -representation. This furthermore defines a derivation tree \mathcal{D}_0 for w_0 (which has 0 occurrences of words u_1, u_2, \dots), which is the same as \mathcal{D} , except that for each pump pair (v, u') , the node v is replaced with the sub-tree of \mathcal{D} with root u' . So as to not split u_i or u_j up, we continue the traversal on u' , which, when it is finished, continues the traversal in the parent of v , having finished with v . Notice that this ensures that each node is in at most one pump pair.

The third property. Consider the word w_0 which has 0 occurrences of words u_1, u_2, \dots . Observe that in derivation tree \mathcal{D}_0 for w_0 , there is at most one occurrence of each non-terminal in each path to the root, since we visited all nodes of \mathcal{D}_0 in our defining traversal and were greedy. Hence, the height is at most T and thus, since the tree is binary, it has at most 2^{T-1} many leaves, which is then a bound on $|w_0| = \sum_{i=1}^{k+1} |s_i|$. Notice that each node of \mathcal{D}_0 , being a subset of \mathcal{D} , is in at most 1 pump pair of w . On the other hand for each pump pair (v, u') of w , we have that u' is a node of \mathcal{D}_0 by construction. Hence, w has at most $2^T - 1$ many pump pairs. Since each pump pair gives rise to at most 2 word $u_i, u_{i'}$, we have $k \leq 2^{T+1} - 2$. □

Sets closed under reachability. Fix an NFA. We say that a set Q' of states of the NFA is *closed under reachability* if for all $q \in Q'$ and $a \in \Sigma$ we have $\delta(q, a) \subseteq Q'$. Clearly, the set of all states is closed under reachability.

Reachability sets. Fix an NFA. Given a state q in the NFA and a word w , let Q_q^w be the set of states reachable upon reading w , starting in q . The set of states $\mathbf{R}(w, q)$ is then the set of states reachable from Q_q^w upon reading any word. For a set Q' and word w , the set $\mathbf{R}(w, Q')$ is $\bigcup_{q \in Q'} \mathbf{R}(w, q)$.

Note the following: For all Q' and w the set $\mathbf{R}(w, Q')$ is closed under reachability. If a set Q' is closed under reachability then $\mathbf{R}(w, Q') \subseteq Q'$ for all w .

We have the following **property of reachability sets**: Fix a word u , a number ℓ , an NFA and a set of states Q' of the NFA, where Q' is closed under reachability. Let u' be a word with ℓ occurrences of u (e.g. u^ℓ). Consider any word w with edit distance strictly less than ℓ from u' . Any run on w , starting in some state of Q' , reaches a state of $\mathbf{R}(u, Q')$. This is because u must be a sub-word of w .

Lemma 19. *Let G be a CFG in CNF with a set of non-terminals of size T and let \mathcal{A}_N be a safety NFA with a set of states Q . The following conditions are equivalent:*

- (i) *the edit distance $\text{ed}(\mathcal{L}(G), \mathcal{L}(\mathcal{A}_N))$ is infinite,*
- (ii) *the edit distance $\text{ed}(\mathcal{L}(G), \mathcal{L}(\mathcal{A}_N))$ exceeds $B := (2^{T+1} - 2) \cdot n + 2^T$, and*

- (iii) there exists a word $w \in \mathcal{L}(G)$, with compact G -decomposition $w = (s_i u_i)_{i=1}^k s_{k+1}$, such that $R(u_k, R(u_{k-1}, R(u_{k-2}, \dots R(u_1, Q) \dots))) = \emptyset$.
- (iv) there exist words u_1, \dots, u_k such that $R(u_k, R(u_{k-1}, R(u_{k-2}, \dots R(u_1, Q) \dots))) = \emptyset$ and for every $\ell > 0$ there exist words s_1, \dots, s_{k+1} such that the word $w_\ell = (s_i u_i^\ell)_{i=1}^k s_{k+1}$ belongs to $\mathcal{L}(G)$.

We use condition (iv) from Lemma 19 later in Section 4.2. Before we proceed with we argue by example that the nested applications of the R function in Lemma 19 is necessary.

The necessity of the recursive applications of the R operator. Consider for instance the alternate requirement that at least one of $R(u_i, Q)$ is empty, for some i . This alternate requirement would not capture that the pushdown language $\{a^n \# b^n \mid n \in \mathbb{N}\}$ has infinite edit distance to the regular language $a^* + b^*$ — for any word in the pushdown language $w = a^n \# b^n$, for some fixed n , a compact G -representation of w is $u_1 = a^n$, $s_2 = \#$ and $u_2 = b^n$ (and the remaining words are empty). But clearly $R(u_1, Q)$ and $R(u_2, Q)$ are not empty since both strings are in the regular language. On the other hand $R(u_2, R(u_1, Q))$ is empty.

Proof. The implication (i) \Rightarrow (ii) is trivial.

We show the implication (ii) \Rightarrow (iii) as follows: Consider a word $w \in \mathcal{L}(G)$ with $ed(w, \mathcal{L}(\mathcal{A}_N)) > B$ and its compact G representation $w = (s_i u_i)_{i=1}^k s_{k+1}$ (which exists due to Lemma 18). We claim that $R(u_k, R(u_{k-1}, R(u_{k-2}, \dots R(u_1, Q) \dots))) = \emptyset$. The argument is by contradiction. Assume that $R(u_k, R(u_{k-1}, R(u_{k-2}, \dots R(u_1, Q) \dots))) \neq \emptyset$ and we will construct a run of \mathcal{A}_N spelling a word w' in $\mathcal{L}(\mathcal{A}_N)$, which has edit distance at most B to w . The description of the run is iteratively in i ; we start with $i = 0$. First, spell out a word s'_i , so that \mathcal{A}_N reaches some state q_i such that there exists a run on u_i . The length of s'_i is at most n . Afterwards follow the run on u_i and go to the next iteration. This run spells the word $w' := (s'_i u_i)_{i=1}^k$. All the choices of q_i 's can be made since $R(u_k, R(u_{k-1}, R(u_{k-2}, \dots R(u_1, Q) \dots))) \neq \emptyset$. Also, since \mathcal{A}_N is a safety automata, this run is accepting. To edit w' into w change each s'_i into s_i and insert s_{k+1} at the end. In the worst case, each s_i is empty except for $i = k + 1$ and in that case it requires $k \cdot n + |w_0| \leq B$ edits for deleting each s'_i and inserting s_{k+1} at the end (in any other case, we would be able to substitute some letters when we change some s'_i into s_i which would make the edit distance smaller). This is a contradiction.

The implication (iii) \Rightarrow (iv) is trivial.

For the implication (iv) \Rightarrow (i) we will argue that for all ℓ , the word $w_\ell \in \mathcal{L}(G)$ requires at least ℓ edits. Consider $w_\ell = (s_i u_i^\ell)_{i=1}^k s_{k+1}$ for some ℓ . Any run on $s_1 u_1^\ell$ (a prefix of w_ℓ) has entered $R(u_1, Q)$ or made at least ℓ edits by the property of reachability sets. Similarly, for any j , any run on $(s_i u_i^\ell)_{i=1}^j$ has either entered $R(u_j, R(u_{j-1}, R(u_{j-2}, \dots R(u_1, Q) \dots)))$ or there has been at least ℓ edits. Since $R(u_k, R(u_{k-1}, R(u_{k-2}, \dots R(u_1, Q) \dots))) = \emptyset$, no run can enter that set and thus there has been at least ℓ edits on w_ℓ . The implication and thus the lemma follows. \square

As a direct consequence of Lemma 19 we have the following.

Theorem 20. (1) For a PDA \mathcal{A}_P and an NFA \mathcal{A}_N we have $ed(\mathcal{L}(\mathcal{A}_P), \mathcal{L}(\mathcal{A}_N))$ is either exponentially bounded in $|\mathcal{A}_P|$ or it is infinite. (2) For $\mathcal{C}_1 \in \{\text{DPDA}, \text{PDA}\}$ we have $\text{FED}(\mathcal{C}_1, \text{NFA})$ is in *ExpTime*

Proof. (1) The equivalence of (i) and (ii) gives a bound on the maximum finite edit distance.

(2) The argument follows from Lemma 6 and (1), i.e., we can check with Lemma 6 TED for k exceeding the bound from (1). \square

4.2 Upper bound for DFA

We show that for $\mathcal{C}_1 \in \{\text{DPDA}, \text{PDA}\}$ the problem $\text{FED}(\mathcal{C}_1, \text{DFA})$ is coNP-complete.

coNP-hardness and attempting to apply known techniques for the upper bound. The lower bound follows directly from the fact that $\text{FED}(\text{DFA}, \text{DFA})$ is coNP-hard [3]. We thus focus on the upper bound. Note that the upper bound was simple for $\text{FED}(\text{DFA}, \text{DFA})$, since the edit distance for such is either polynomial or infinite and there is a polynomial length witness in case it is infinite. Hence, one just guess the polynomial sized witness w and runs a polynomial time algorithm for $ed(w, \text{DFA})$ and the result follows. Doing the similar thing for $\text{FED}(\text{PDA}, \text{DFA})$

would give a NExpTime upper-bound, since the word we need to guess might be of exponential length (thus the above ExpTime upper bound for FED(PDA, NFA) is better). To give our algorithm, we will first define extended reachability sets and give a key proposition.

Closed under concatenation and extended reachability sets. A language L is said to be *closed under concatenation* if for all $w_1, w_2 \in L$ we have $w_1 w_2 \in L$. Note that $\mathcal{L}(G, A, D)$, for any non-terminal A and direction D , is always closed under concatenation.

We extend reachability sets as follows: Let L be a context-free language closed under concatenation, let \mathcal{A}_D be a DFA and let Q' be a subset of the states of \mathcal{A}_D . We define $R(L, Q')$ as the intersection $\bigcap_{w \in L} R(w, Q')$. Observe that for every L there exists a finite subset $W \subseteq L$ such that $\bigcap_{w \in W} R(w, Q') = R(L, Q')$.

Remark 21. If Q' is closed under reachability, then for any set of words $W = \{w_1, w_2, \dots, w_k\} \subseteq L$ such that $\bigcap_{w \in W} R(w, Q') = R(L, Q')$, we have that $w' = w_1 w_2 \dots w_k \in L$ and $R(w', Q') = R(L, Q')$. The latter comes from the fact that for any word w'' and set Q'' closed under reachability, we have that $R(s_1 w'' s_2, Q'') \subseteq R(w'', Q'')$ for all s_1 and s_2 .

Also, observe that we have the following facts about R , from the definition of R :

1. For any $Q'' \subseteq Q'$ and word w we have that $R(w, Q'') \subseteq R(w, Q')$.
2. For any language L , any Q' and word $w \in L$, we have that $R(L, Q') \subseteq R(w, Q')$.

The following proposition is a key to our coNP-algorithm.

Proposition 22. For any k , any sequence of languages L_1, \dots, L_k and any word $w_i \in L_i$ for each i , we have

$$R(L_k, \dots, R(L_1, Q) \dots) \subseteq R(w_k, \dots, R(w_1, Q) \dots) .$$

Also, if each L_i is closed under concatenation, then there exist words $w'_i \in L_i$ for each i , such that

$$R(L_k, \dots, R(L_1, Q) \dots) = R(w'_k, \dots, R(w'_1, Q) \dots)$$

Proof. The proposition follows from Remark 21 and simple induction. □

coNP-upper bound algorithm. Our coNP-algorithm InfEdsSeq deciding whether the edit distance is finite works as follows:

1. Guess a sequence $s = a_{D_1}^{A_1} a_{D_2}^{A_2} \dots a_{D_k}^{A_k}$, for some k .
2. return “no” if s is such that (1) s is realizable; and (2)

$$R(\mathcal{L}(G, A_k, D_k), R(\mathcal{L}(G, A_{k-1}, D_{k-1}), \dots R(\mathcal{L}(G, A_1, D_1), Q) \dots)) = \emptyset .$$

3. otherwise return yes.

Requirements for InfEdsSeq to be in coNP. For InfEdsSeq to be in coNP, we need to give the following:

1. A polynomial bound on k (so that s is a polynomial sized witness). The bound will be given in Lemma 24.
2. A polynomial time algorithm to decide whether a sequence $s = a_{D_1}^{A_1} a_{D_2}^{A_2} \dots a_{D_k}^{A_k}$ is realizable. The algorithm will be given in Lemma 25.
3. A polynomial time algorithm for computing $R(\mathcal{L}(G, A, D), Q')$ for any CFG G , any non-terminal A , any direction D and any set Q' closed under reachability. This will allow us to decide, given a realizable sequence $s = a_{D_1}^{A_1} a_{D_2}^{A_2} \dots a_{D_k}^{A_k}$, whether

$$R(\mathcal{L}(G, A_k, D_k), R(\mathcal{L}(G, A_{k-1}, D_{k-1}), \dots R(\mathcal{L}(G, A_1, D_1), Q) \dots)) = \emptyset ,$$

by evaluating the expression on the left-hand side inside-out. The algorithm for computing $R(\mathcal{L}(G, A, D), Q')$ will be given in Corollary 27.

We will first argue that the algorithm is correct.

Lemma 23. *The algorithm InfEdsSeq is correct.*

Proof. To argue that the algorithm is correct, we just need to argue that a sequence with properties (1) and (2) exists if and only if the edit distance is infinite.

Such a sequence implies infinite edit distance. According to Proposition 22, such a sequence indicates that there are words $w_i \in \mathcal{L}(G, A_i, D_i)$ for each i , such that

$$R(w_k, \dots, R(w_1, Q) \dots)) = \emptyset.$$

For all i , since $\mathcal{L}(G, A_i, D_i)$ is closed under concatenation, we also have that $w_i^\ell \in \mathcal{L}(G, A_i, D_i)$ for all $\ell > 0$. Thus, by Lemma 17, there exist words s_1, \dots, s_k, s_{k+1} such that $s_1 w_1^\ell \dots s_k w_k^\ell s_{k+1}$ belongs to $\mathcal{L}(G)$. Hence, item (iv) of Lemma 19 is satisfied and we get that the edit distance is infinite.

Infinite edit distance implies the existence of such a sequence. When the edit distance is infinite, then, according to item (iii) of Lemma 19 there exists a word $w \in \mathcal{L}(G)$, with compact G -decomposition $w = (s_i u_i)_{i=1}^k s_{k+1}$, such that $R(u_k, R(u_{k-1}, R(u_{k-2}, \dots, R(u_1, Q) \dots))) = \emptyset$. By definition of compact G -decomposition, every u_i from the decomposition is associated with a terminal $a_{D_i}^{A_i}$, such that the sequence $a_{D_1}^{A_1} \dots a_{D_k}^{A_k}$ is realizable (satisfying property (1)) and $u_i \in \mathcal{L}(G, A_i, D_i)$ for each i . By Proposition 22 we then have that

$$R(\mathcal{L}(G, A_k, D_k), R(\mathcal{L}(G, A_{k-1}, D_{k-1}), \dots, R(\mathcal{L}(G, A_1, D_1), Q) \dots)) = \emptyset,$$

(satisfying property (1)). Thus such a sequence always exists and the lemma follows. \square

Next, we will give the bounds and algorithms to show that InfEdsSeq is in coNP. First the bound on k .

Lemma 24. *Let G be a CFG and let \mathcal{A}_D be a safety DFA with a set of states Q . The following conditions are equivalent:*

- (i) *the edit distance $ed(\mathcal{L}(G), \mathcal{L}(\mathcal{A}_D))$ is infinite.*
- (ii) *there exists a realizable sequence $(a_{D_i}^{A_i})_{i=1}^m$ with $m \leq |Q|$ such that*

$$R(\mathcal{L}(G, A_m, D_m), R(\mathcal{L}(G, A_{m-1}, D_{m-1}), \dots, R(\mathcal{L}(G, A_1, D_1), Q) \dots)) = \emptyset.$$

Proof. (i) implies (ii). Assume that $ed(\mathcal{L}(G), \mathcal{L}(\mathcal{A}_D))$ is infinite. By Lemma 19, there exists a word $w \in \mathcal{L}(G)$, with compact G -decomposition $w = (s_i u_i)_{i=1}^k s_{k+1}$, such that $R(u_k, R(u_{k-1}, \dots, R(u_1, Q) \dots)) = \emptyset$. Observe that k can be exponential. We claim that we can pick from u_1, \dots, u_k a sub-sequence of polynomial length in $|Q|$ for which the reachable set of states is empty as well. Indeed, the sequence $s = R(u_1, Q), R(u_2, R(u_1, Q)), \dots$ is weakly decreasing with respect to the set inclusion (i.e. if a state is not in $s[i]$, then, it cannot be in $s[j]$ for $j \geq i$, because R is closed under reachability). We select from $1, \dots, k$ indices i on which the sequence $R(u_1, Q), R(u_2, R(u_1, Q)), \dots$ strictly decreases and denote the resulting sub-sequence by α . Then,

$$R(u_{\alpha(m)}, R(u_{\alpha(m-1)}, \dots, R(u_{\alpha(1)}, Q) \dots)) = \emptyset.$$

There are at most $|Q|$ such indices, therefore $|\alpha| = m \leq |Q|$. Using Proposition 22, since $u_{\alpha(i)} \in \mathcal{L}(G, A_{\alpha(i)}, D_{\alpha(i)})$ by compact G -decomposition, we get that:

$$R(\mathcal{L}(G, A_{\alpha(m)}, D_{\alpha(m)}), \dots, R(\mathcal{L}(G, A_{\alpha(1)}, D_{\alpha(1)}), Q) \dots) \subseteq R(u_{\alpha(m)}, \dots, R(u_{\alpha(1)}, Q) \dots) = \emptyset.$$

(ii) implies (i). Assume that condition (ii) holds. Then, the algorithm InfEdsSeq returns YES, and its correctness (Lemma 23) implies (i). \square

Next we will describe the algorithm deciding whether a sequence is realizable.

Lemma 25. *Let G be a CFG. We can decide in polynomial time whether a given sequence $s = a_{D_1}^{A_1} \dots a_{D_k}^{A_k}$ is realizable.*

Proof. Consider grammar \bar{G} associated with G . We convert \bar{G} to CNF and add productions $A \rightarrow \epsilon$ for every non-terminal. Let the resulting CFG be G' . Observe that G' derives a word of terminals and non-terminals u if and only if \bar{G} derives a word u' such that u is a subsequence of u' . Thus, $(A_1, D_1), \dots, (A_k, D_k)$ is realizable if and only if $A_1^{D_1} \dots A_k^{D_k}$ is derivable by G' . Since G' has polynomial size in G , we can check whether a word is derivable in G' in polynomial time. \square

Finally, we present the algorithm that computes $R(\mathcal{L}(G, A, D), Q')$. The result will follow as a corollary of the following lemma.

Lemma 26. *Given a CFG G , such that $\mathcal{L}(G)$ is closed under concatenation, a DFA \mathcal{A}_D with a set of states $!$ and a set of states $Q' \subseteq Q$ closed under reachability, the set $R(\mathcal{L}(G), Q')$ is computable in polynomial time.*

Proof. Given a set of states $S \subseteq Q$, we define $\text{Reach}(S)$ as the set of states reachable from Q' in \mathcal{A}_D . We can divide Q' into strongly connected components (SCCs). We say that an SCC C is *recurrent* if \mathcal{A}_D can stay in C upon reading any word from $\mathcal{L}(G)$.

We claim that $R(\mathcal{L}(G), Q')$ is the set of states Q^* reachable from all recurrent SCCs in Q' . Clearly, Q^* is closed under reachability.

- We will first argue that $Q^* \subseteq R(\mathcal{L}(G), Q')$. First, for every recurrent SCC C and every word $w \in \mathcal{L}(G)$, there is a state $s' \in C$ such that $R(w, s') \in C$. Therefore, $C \subseteq R(w, s')$. By Remark 21, it follows that $C \subseteq R(\mathcal{L}(G), Q')$ and $Q^* \subseteq R(\mathcal{L}(G), Q')$.
- We will next argue that $R(\mathcal{L}(G), Q') \subseteq Q^*$. Observe that for every state s in a non-recurrent SCC C there exists a word $w_s \in \mathcal{L}(G)$ that forces \mathcal{A}_D to leave C , i.e., $R(w_s, s) \cap C = \emptyset$. Thus, $|R(w_s, C) \cap C| < |C|$. It follows that we can remove states from $R(w_s, C) \cap C$ one by one by concatenating words w_s to obtain a word w_C such that $R(w_C, C) \cap C = \emptyset$. Since $\mathcal{L}(G)$ is closed under concatenation, the word w_C belongs to $\mathcal{L}(G)$.

Let C_1, C_2, \dots, C_ℓ be the SCCs in Q' not in Q^* (and thus non-recurrent) ordered topologically. Let the word w_T be the word $w_T = w_{C_1} w_{C_2} \dots w_{C_\ell}$. Observe that $w_T \in \mathcal{L}(G)$. We have $R(\mathcal{L}(G), Q') \subseteq R(w_T, Q')$ by Remark 21 and we argue that $R(w_T, Q') \subseteq Q^*$.

Any run starting in Q^* will end in Q^* , since Q^* is closed under reachability. Observe that $R(w_{C_1}, C_1 \cup \dots \cup C_\ell)$ does not contain C_1 as $R(w_{C_1}, C_1) \cap C_1 = \emptyset$ and due to topological order C_1 is not reachable from C_2, \dots, C_ℓ . Thus, by induction reasoning we have $R(w_{C_1}, R(w_{C_2}, \dots, R(w_{C_\ell}, C_\ell) \dots))$ does not contain C_1, \dots, C_ℓ . Observe that $R(w_T, C_1 \cup \dots \cup C_\ell) \subseteq R(w_{C_1}, R(w_{C_2}, \dots, R(w_{C_\ell}, C_\ell) \dots))$, and hence $R(w_T, C_1 \cup \dots \cup C_\ell) \subseteq Q^*$.

Given a SCC C and a state $s \in C$, let the automaton $\mathcal{A}_D^{C,s}$ be \mathcal{A}_D restricted to C and with start state s . Observe that a SCC C is recurrent if and only if there is a state $s \in C$ such that $\mathcal{L}(G) \subseteq \mathcal{L}(\mathcal{A}_D^{C,s})$. We can then easily test if a SCC is recurrent by trying each possibility for $s \in C$ and testing if $\mathcal{L}(G) \subseteq \mathcal{L}(\mathcal{A}_D^{C,s})$. This can be done in polynomial time since language inclusion of a CFG in a DFA can be tested in polynomial time.

Thus our algorithm is as follows: Compute the set $\{C_1, \dots, C_\ell\}$ of SCCs in Q' . For each i test if C_i is recurrent and let $\{C'_1, \dots, C'_{\ell'}\}$ be the recurrent SCCs in Q' . Return $\bigcup_{i=1}^{\ell'} \text{Reach}(C'_i)$. \square

We next get the wanted corollary.

Corollary 27. *Given a CFG G , a non-terminal A , a direction D , a DFA \mathcal{A}_D and a set of states Q' of \mathcal{A}_D closed under reachability, the set $R(\mathcal{L}(G, A, D), Q')$ is computable in polynomial time.*

Proof. The proof follows from Lemma 16 and Lemma 26, using that $\mathcal{L}(G, A, D)$ is closed under concatenation. \square

Lemma 28. *Given a context-free grammar G and a safety DFA \mathcal{A}_D the algorithm *InfEdsSeq* can be implemented in *coNP* and correctly decides whether $\text{ed}(\mathcal{L}(\mathcal{A}_P), \mathcal{L}(\mathcal{A}_D))$ is finite. Moreover, if \mathcal{A}_D is of constant size then *InfEdsSeq* does not need non-determinism (and thus uses polynomial time only).*

Proof. The correctness comes from Lemma 23. The complexity comes from Lemma 24, Lemma 25 and Corollary 27. Note that, in case the DFA is of constant size, then k is bounded by a constant, according to Lemma 24, and thus there are only a polynomial number of candidates for s and hence all can be checked using polynomial time in total. \square

Theorem 29. For $\mathcal{C}_1 \in \{\text{DPDA}, \text{PDA}\}$ we have $\text{FED}(\mathcal{C}_1, \text{DFA})$ is *coNP-complete*.

Proof. First, we discuss containment of $\text{FED}(\text{PDA}, \text{DFA})$ in *coNP*. Consider a PDA \mathcal{A}_P and a DFA \mathcal{A}_D . We can transform \mathcal{A}_P to a context-free grammar G with $\mathcal{L}(\mathcal{A}_P) = \mathcal{L}(G)$ in polynomial time. Also, we can transform \mathcal{A}_D to a safety DFA \mathcal{A}'_D recognizing the language $\text{prefix}(\mathcal{L}(\mathcal{A}_D))$. Due to Lemma 14, we have $\text{ed}(\mathcal{L}(G), \mathcal{A}_D)$ is finite if and only if $\text{ed}(\mathcal{L}(G), \mathcal{A}'_D)$ is finite. By Lemma 28 we can decide whether $\text{ed}(\mathcal{L}(G), \mathcal{A}'_D)$ is finite in *coNP*. Hence, $\text{FED}(\text{PDA}, \text{DFA})$ and $\text{FED}(\text{DPDA}, \text{DFA})$ are in *coNP*.

It has been shown in [3] that $\text{FED}(\text{DFA}, \text{DFA})$ is *coNP-hard*, therefore $\text{FED}(\text{PDA}, \text{DFA})$ and $\text{FED}(\text{DPDA}, \text{DFA})$ are *coNP-hard*. \square

4.3 Lower bound

We have shown the exponential upper bound on the edit distance if it is finite. As mentioned in the introduction, it is easy to define a family of context free grammars only accepting an exponential length word, using repeated doubling and thus the edit distance can be exponential between DPDAs and DFAs. We can also show that the inclusion problem reduces to the finite edit distance problem $\text{FED}(\text{DPDA}, \text{NFA})$ and get the following lemma.

Lemma 30. $\text{FED}(\text{DPDA}, \text{NFA})$ is *ExpTime-hard*.

Proof. We show that the inclusion problem of DPDA in NFA, which is *ExpTime-hard* by Lemma 8 reduces to $\text{FED}(\text{DPDA}, \text{NFA})$. Consider a DPDA \mathcal{A}_P and an NFA \mathcal{A}_N . We define $\widehat{\mathcal{L}} = \{\#w_1\# \dots \#w_k\# : k \in \mathbb{N}, w_1, \dots, w_k \in \mathcal{L}\}$. Observe that either $\widehat{\mathcal{L}}_1 \subseteq \widehat{\mathcal{L}}_2$ or $\text{ed}(\widehat{\mathcal{L}}_1, \widehat{\mathcal{L}}_2) = \infty$. Therefore, $\text{ed}(\widehat{\mathcal{L}}_1, \widehat{\mathcal{L}}_2) < \infty$ if and only if $\mathcal{L}_1 \subseteq \mathcal{L}_2$. In particular, $\mathcal{L}(\mathcal{A}_P) \subseteq \mathcal{L}(\mathcal{A}_N)$ if and only if $\text{ed}(\widehat{\mathcal{L}}(\mathcal{A}_P), \widehat{\mathcal{L}}(\mathcal{A}_N)) < \infty$. Observe that in polynomial time we can transform \mathcal{A}_P (resp., \mathcal{A}_N) to a DPDA $\widehat{\mathcal{A}}_P$ (resp., an NFA $\widehat{\mathcal{A}}_N$) recognizing $\widehat{\mathcal{L}}(\mathcal{A}_P)$ (resp., $\widehat{\mathcal{L}}(\mathcal{A}_N)$). It suffices to add transitions from all final states to all initial states with the letter $\#$, i.e., $\{(q, \#, s) : q \in F, s \in S\}$ for NFA (resp., $\{(q, \#, \perp, s) : q \in F, s \in S\}$ for DPDA). For DPDA the additional transitions are possible only with empty stack. \square

5 Edit distance to PDA

Observe that the threshold distance problem from DFA to PDA with the fixed threshold 0 and a fixed DFA recognizing Σ^* coincides with the universality problem for PDA. Hence, the universality problem for PDA, which is undecidable, reduces to $\text{TED}(\text{DFA}, \text{PDA})$. The universality problem for PDA reduces to $\text{FED}(\text{DFA}, \text{PDA})$ as well by the same argument as in Lemma 30. Finally, we can reduce the inclusion problem from DPDA in DPDA, which is undecidable, to $\text{TED}(\text{DPDA}, \text{DPDA})$ (resp., $\text{FED}(\text{DPDA}, \text{DPDA})$). Again, we can use the same construction as in Lemma 30. In conclusion, we have the following proposition.

Proposition 31. (1) For every class $\mathcal{C} \in \{\text{DFA}, \text{NFA}, \text{DPDA}, \text{PDA}\}$, the problems $\text{TED}(\mathcal{C}, \text{PDA})$ and $\text{FED}(\mathcal{C}, \text{PDA})$ are undecidable. (2) For every class $\mathcal{C} \in \{\text{DPDA}, \text{PDA}\}$, the problem $\text{FED}(\mathcal{C}, \text{DPDA})$ is undecidable.

The results in (1) of Proposition 31 are obtained by reduction from the universality problem for PDA. However, the universality problem for DPDA is decidable. Still we show that $\text{TED}(\text{DFA}, \text{DPDA})$ is undecidable. The overall argument is similar to the one in Section 3.2. First, we define nearly-deterministic PDA, a pushdown counterpart of nearly-deterministic NFA.

Definition 32. A PDA $\mathcal{A} = (\Sigma, \Gamma, Q, S, \delta, F)$ is nearly-deterministic if $|S| = 1$ and $\delta = \delta_1 \cup \delta_2$, where δ_1 is a function and for every accepting run, the automaton takes a transition from δ_2 exactly once.

By carefully reviewing the standard reduction of the halting problem for Turing machines to the universality problem for pushdown automata [14], we observe that the PDA that appear as the product of the reduction are nearly-deterministic.

Lemma 33. *The problem, given a nearly-deterministic PDA \mathcal{A}_P , decide whether $\mathcal{L}(\mathcal{A}_P) = \Sigma^*$, is undecidable.*

Using the same construction as in Lemma 10 we show a reduction of the universality problem for nearly-deterministic PDA to $\text{TED}(\text{DFA}, \text{DPDA})$.

Proposition 34. *For every class $\mathcal{C} \in \{\text{DFA}, \text{NFA}, \text{DPDA}, \text{PDA}\}$, the problem $\text{TED}(\mathcal{C}, \text{DPDA})$ is undecidable.*

Proof. We show that $\text{TED}(\text{DFA}, \text{DPDA})$ (resp., $\text{FED}(\text{DFA}, \text{PDA})$) is undecidable as it implies undecidability of the rest of the problems. The same construction as in the proof of Lemma 10 shows a reduction of the universality problem for nearly-deterministic PDA, which is undecidable by Lemma 33, to $\text{TED}(\text{DFA}, \text{DPDA})$. \square

We presented the complete decidability picture for the problems $\text{TED}(\mathcal{C}_1, \mathcal{C}_2)$, for $\mathcal{C}_1 \in \{\text{DFA}, \text{NFA}, \text{DPDA}, \text{PDA}\}$ and $\mathcal{C}_2 \in \{\text{DPDA}, \text{PDA}\}$. To complete the characterization of the problems $\text{FED}(\mathcal{C}_1, \mathcal{C}_2)$, with respect to their decidability, we still need to settle the decidability (and complexity) status of $\text{FED}(\text{DFA}, \text{DPDA})$. We leave it as an open problem, but conjecture that it is coNP -complete.

Conjecture 35. *$\text{FED}(\text{DFA}, \text{DPDA})$ is coNP -complete.*

6 Conclusions

In this work we consider the edit distance problem for PDA and its subclasses and present a complete decidability and complexity picture for the TED problem. We leave some open conjectures about the parametrized complexity of the TED problem, and the complexity of FED problem when the target is a DPDA. Moreover, one can study the edit distance for other classes of languages between regular languages and context-free languages such as visibly pushdown automata.

While in this work we count the number of edit operations, a different notion is to measure the average number of edit operations. The average-based measure is undecidable in many cases even for finite automata, and in cases when it is decidable reduces to mean-payoff games on graphs [4]. Since mean-payoff games on pushdown graphs are undecidable [10], most of the problems related to the edit distance question for average measure for DPDA and PDA are likely to be undecidable.

Acknowledgements. We wanted to thank the anonymous reviewers for their thorough and helpful reviews, which help us to improve this paper.

References

- [1] A. Aho and T. Peterson. A minimum distance error-correcting parser for context-free languages. *SIAM J. of Computing*, 1:305–312, 1972.
- [2] Michael Benedikt, Gabriele Puppis, and Cristian Riveros. Regular repair of specifications. In *LICS'11*, pages 335–344, 2011.
- [3] Michael Benedikt, Gabriele Puppis, and Cristian Riveros. Bounded repairability of word languages. *J. Comput. Syst. Sci.*, 79(8):1302–1321, 2013.
- [4] Michael Benedikt, Gabriele Puppis, and Cristian Riveros. The per-character cost of repairing word languages. *Theor. Comput. Sci.*, 539:38–67, 2014.
- [5] Ashok K. Chandra, Dexter C. Kozen, and Larry J. Stockmeyer. Alternation. *J. ACM*, 28(1):114–133, January 1981.

- [6] Krishnendu Chatterjee, Laurent Doyen, and Thomas A. Henzinger. Quantitative languages. *ACM Trans. Comput. Log.*, 11(4), 2010.
- [7] Krishnendu Chatterjee, Thomas A. Henzinger, Rasmus Ibsen-Jensen, and Jan Otop. Edit distance for pushdown automata. In *Automata, Languages, and Programming - 42nd International Colloquium, ICALP 2015, Kyoto, Japan, July 6-10, 2015, Proceedings, Part II*, pages 121–133, 2015.
- [8] Krishnendu Chatterjee, Thomas A. Henzinger, and Jan Otop. Nested weighted automata. *CoRR*, abs/1504.06117, 2015. To appear at LICS 2015.
- [9] Krishnendu Chatterjee, Rasmus Ibsen-Jensen, and Rupak Majumdar. Edit distance for timed automata. In *HSCC’14*, pages 303–312, 2014.
- [10] Krishnendu Chatterjee and Yaron Velner. Mean-payoff pushdown games. In *LICS*, pages 195–204, 2012.
- [11] Noam Chomsky. On certain formal properties of grammars. *Information and Control*, 2(2):137 – 167, 1959.
- [12] Paweł Gawrychowski. Faster algorithm for computing the edit distance between slp-compressed strings. In *SPIRE’12*, pages 229–236, 2012.
- [13] Thomas A. Henzinger and Jan Otop. From model checking to model measuring. In *CONCUR’13*, pages 273–287, 2013.
- [14] John E. Hopcroft and Jefferey D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Adison-Wesley Publishing Company, Reading, Massachusetts, USA, 1979.
- [15] R.M. Karp. Mapping the genome: some combinatorial problems arising in molecular biology. In *STOC 93*, pages 278–285. ACM, 1993.
- [16] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.
- [17] Yuri Lifshits. Processing compressed texts: A tractability border. In *Combinatorial Pattern Matching*, pages 228–240. Springer, 2007.
- [18] M. Mohri. Edit-distance of weighted automata: general definitions and algorithms. *Intl. J. of Foundations of Comp. Sci.*, 14:957–982, 2003.
- [19] T. Okuda, E. Tanaka, and T. Kasai. A method for the correction of garbled words based on the levenshtein metric. *IEEE Trans. Comput.*, 25:172–178, 1976.
- [20] G. Pighizzini. How hard is computing the edit distance? *Information and Computation*, 165:1–13, 2001.
- [21] Barna Saha. The dyck language edit distance problem in near-linear time. In *FOCS’14*, pages 611–620, 2014.
- [22] Robert A. Wagner and Michael J. Fischer. The string-to-string correction problem. *J. ACM*, pages 168–173, 1974.